

Open Science Grid

We propose that the DOE and the NSF endorse a roadmap for the U.S. to build a national grid infrastructure for science: the Open Science Grid.

We propose an aggressive program of work to federate many of the currently disjoint grid resources at labs and universities into a single scalable, engineered, and managed grid.

Starting with the U.S. LHC grid resources a fully functional and production quality grid will be built and operated, that supports Peta-scale operations, and that extends internationally to create a global grid for LHC science. This grid will immediately serve as a backbone to merge grid computing efforts of other experiments in particle and nuclear physics, and can rapidly be extended to other science communities.

The Open Science Grid will ensure that current and future particle and nuclear physics experiments benefit maximally from the large computing, data storage, network and grid investments already made and planned for. The Open Science Grid will ensure that the U.S. plays a leading role in defining and operating the global grid infrastructure needed for large-scale collaborative and international scientific research. The Open Science Grid will provide a set of services that can be enriched as new science areas choose to join and federate their resources. For the first time combined computing resources at several DOE labs and at dozens of universities will effectively become a single national computing infrastructure for science, the Open Science Grid.

The Open Science Grid will provide opportunities for educators and students to participate in building and exploiting this grid infrastructure and opportunities for developing and training a scientific and technical workforce. It has potential to transform the integration of education and research at all levels.

Roadmap to the Open Science Grid

The LHC is an exemplar global science project supported by a partnership of DOE and NSF. It is exemplary not only as a worldwide international collaboration in large-scale science, but also for driving the use of grids. The LHC experiments are driving grid technology in order to be able to share the costs and burdens of the immense computing and storage resources needed and, more importantly, to enable the scientists working on the experiments to fully participate in the science from the far corners of the globe.

U.S.-developed grid middleware and emerging grid services¹ are at the base of all the LHC grid work, and up to now the U.S. prototypical and experiment-specific grids² have led the way with demonstrations of how the global computing and data grid system can be built.

¹ Globus (<http://globus.org>), Condor (<http://cs.wisc.edu/condor/>), and others

² Through the experiments' test-beds and the U.S. Physics Grid projects PPDG (<http://ppdg.net/>), iVDGL (<http://ivdgl.org/>), and GriPhyN (<http://griphyn.org/>)

Much of the LHC grid infrastructure in Europe will be provided by a combination of CERN central resources and a consortium of European centers that propose³ to federate some of their computing resources in a grid for e-science in Europe. Centers in the U.S. and other parts of the world will federate with this European grid infrastructure in order to provide the global computing grid for LHC science.

It is now time for the U.S. to also federate its LHC computing resources and in doing so to continue to lead the efforts in constructing a global grid for LHC science. We propose to provide and operate these resources at the national laboratories and universities as the seed for the Open Science Grid.

The work of federating U.S. computing resources into a scalable, well-engineered and managed production quality grid for LHC science will allow us to build the Open Science Grid as a new national grid infrastructure serving thousands of scientists.

³ Proposal "Enabling Grids for E-Science and Industry in Europe, <http://egee-ci.web.cern.ch/egee-ci/New/Home.htm>, a \$33M€ proposal to the European Commission, over 2 years, starting in 2004

In building this new and bold computational infrastructure for science, the U.S. will be in a global leadership position for computational and data intensive science and the Open Science Grid will benefit a large number of stakeholders in fundamental and new ways. This will have a broad impact on research in numerous fields and shape the future use of grids in industry.

As each experiment-specific set of applications, grid services and computing fabrics are incorporated into the Open Science Grid we will enrich and extend this national grid. The initial focus will be physics, serving the needs of existing and next generation programs. The Open Science Grid that will emerge will provide an infrastructure for other sciences to incorporate, forming the basis for a global computing grid for science.

The Open Science Grid creates a seamless environment for science collaborations and applications: a virtual computing service and a ubiquitous responsive work environment for scientists. It enables individuals and groups of researchers at universities throughout the U.S. to become full participants in a new generation of worldwide science projects, and allows them to collaborate with science groups around the world, lowering the entry barrier for smaller communities to use grid computing.

The Open Science Grid will benefit a range of organizations that are providing computing for science: university computing centers, national laboratories, and the private sector. It will allow them to provide and manage their facilities across their broader program and to capitalize on economies made possible by sharing expertise and support.

The Open Science Grid will build on R&D grid efforts^{1, 2, 3}, using services provided by existing grid organizations⁴ and facilities provided by allied consortia⁵.

Work Plan

In the roadmap to build the Open Science Grid we envisage multiple phases. The Open Science Grid will grow by adding experiments and

science communities, thus expanding the number of participating sites, and by collaborating with other grid computing infrastructures like the TeraGrid and the DOE Science Grid.

The initial phase of the Open Science Grid is to federate the LHC physics applications' grid services and computing resources in the U.S. into a global grid system, engineered and managed to serve the needs of the LHC scientific program. National laboratories and universities participating in the U.S. LHC software and computing efforts will form the initial sites with special roles for the U.S. LHC Tier-1 centers at BNL and Fermilab.

In the next phase, applications from other physics communities, specifically RunII experiments at Fermilab, RHIC experiments at BNL, and BaBar at SLAC, will move their resources and applications to the Open Science Grid. Other experiments and communities, like our PPDG and iVDGL partners, will join and further extend the Open Science Grid. In subsequent phases, other non-physics science applications will be included.

Each application will bring dedicated computing resources to be federated with the Open Science Grid. We expect that the initial costs of integrating a new application will be partially offset through economies of scale, although considerable up-front investment will be necessary to reap the eventual benefits and cost savings of a robust shared national grid infrastructure.

The construction of the Open Science Grid requires the following work elements:

1. **Management and technical oversight.**
2. **Engineering and quality assurance:** to supply the architectural foundation and engineering for the entire system.
3. **International coordination:** to assure interoperability with non-U.S. grids (e.g. LHC Computing Grid).
4. **Education and Outreach:** to provide opportunities for students to witness and participate in building this emerging national grid infrastructure.
5. **Application Integration:** an iterative program of work to integrate each specific application's grid services and

⁴ e.g. the DOE Science Grid
(<http://doesciencegrid.org>)

⁵ e. g. the TeraGrid project
(<http://www.teragrid.org/>)

computing fabric. *An initial such application integration cycle is already underway for U.S. LHC, as a demonstration of the viability of this approach.*⁶

The measure of success for this program will be performance indicators to be defined for each of the applications, as well as the ability to run on resources not “owned” by the Application thus making effective use of the Open Science Grid.

Budget

We estimate that a budget of between \$40M and \$50M spread over 4 years will be required to construct the core Open Science Grid and to provide some funding for studies and functional demonstrators for several of the candidate applications that would join in the second and subsequent phases. This is a remarkably small investment to achieve such an ambitious national grid infrastructure and several important factors must be noted.

- *Investments, primarily at national labs, for computing and services in support of running physics experiments exceed \$30M per year in the U.S. plus significant contributions from European and other foreign collaborators. European computing contributions to U.S. experiments are expected to diminish rapidly close to LHC turn-on.*
- *Investments in LHC computing and services in the U.S. are planned to reach \$30M to \$35M per year in 2006.*
- *Running experiments will require additional resources and encouragement to move to the Open Science Grid, without disrupting their physics research. Institutions that serve both the LHC and these experiments will be in a position to provide part of this help.*
- *European investments in a European Grid infrastructure for e-science are expected to exceed \$16M/year in the next 2 years.*³

- *Some application areas may be able to benefit from the Open Science Grid at almost no additional cost, while others may require considerable work to adapt both the Open Science Grid and their applications and services.*

To assure success there must be a constant effort of approximately 15 FTEs applied to the work elements 1-4. Management (4 FTEs), Engineering (4 FTEs), International Coordination (4 FTEs) and Education and Outreach (3 FTEs). **(\$2M/year total)**

In order to migrate the experiment-specific grid services to the Open Science Grid some core servers and services owned and operated by the federation, rather than simply contributed by one of the partners, will have to be put in place and operated as a robust round the clock service **(\$2M/year).**

The initial phase work of federating the LHC resources and migrating both CMS and ATLAS applications to a common U.S. grid is estimated to require approximately **\$4 M in the first year** in development costs and subsequently between **\$6M and \$7.5M per year** in ongoing integration and operations costs.

In parallel, opportunities to consolidate services and integrate non-LHC experiments and resources should be studied and demonstrated **(\$750K/year).**

The additional funding necessary to fully migrate a particular experiment or application area in phases two and beyond will then be better understood and detailed proposals for this will have to be developed.

⁶ The Grid2003 functional Grid demonstration (<http://www.ivdgl.org/grid3/>) is moving the initial production environments of U.S. LHC onto a Grid based on common middleware, the VDT (<http://lsc-group.phys.uwm.edu/vdt/>).